HELP NGO Dataset CaseStudy

By:
Manoj Kumar

Problem Statement

- The main aim of this case study is to categorise the countries using some socio-economic and health factors that determine the overall development of the country
- We need to find the countries by categorizing into clusters based on **[gdpp, child_mort and income]**

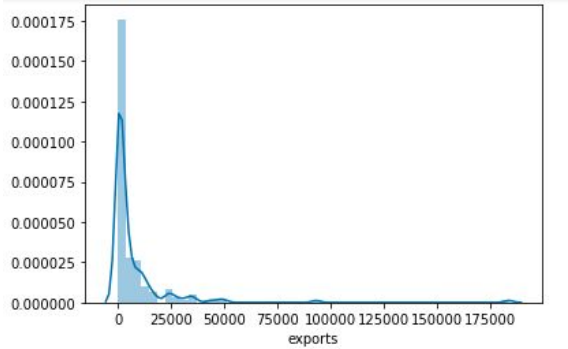
1. Data Quality Check

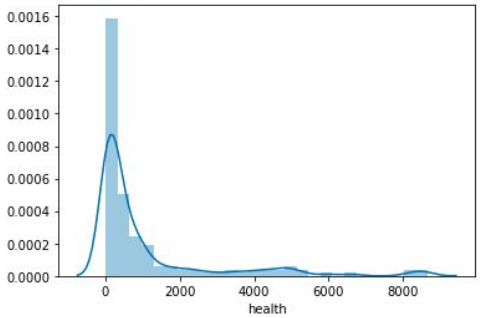
In the given data, columns exports, health, imports are given in percentage terms we need to convert the above columns into numerical values.
Resultant Output is:

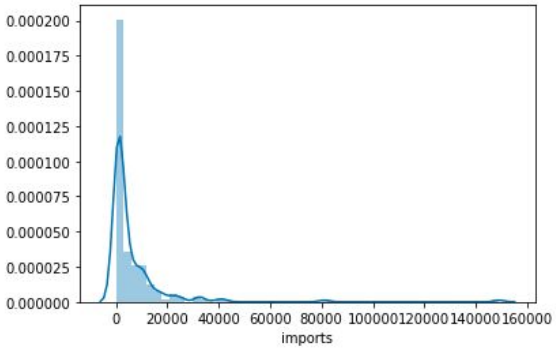
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553
1	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090
2	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460
3	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530
4	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200

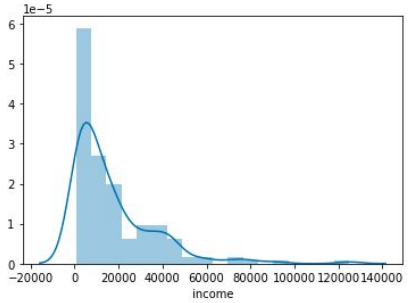
2. EDA: Univariate and Bivariate Analysis

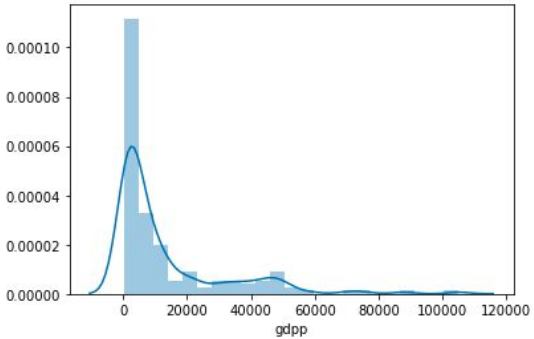
Respective distplots for all the numerical variables in the given set is shown below;

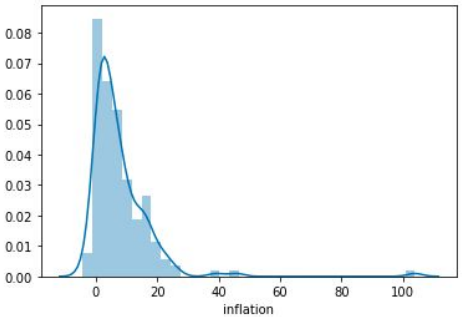


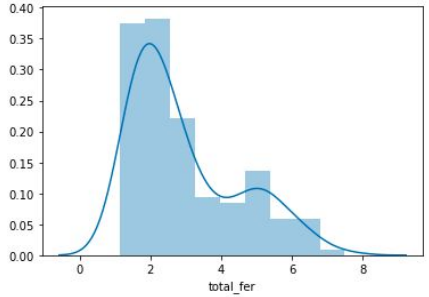


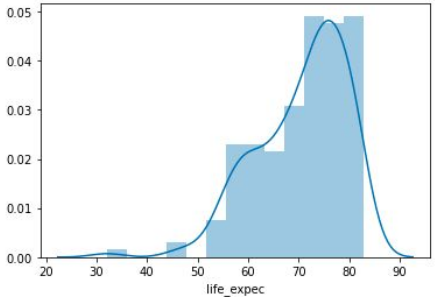

0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
0 20000 40000 60000 80000 100000 120000 140000 160000
imports


1e−5
6
5
4
3
2
1
0
−20000 0 20000 40000 60000 80000 100000 120000 140000
income

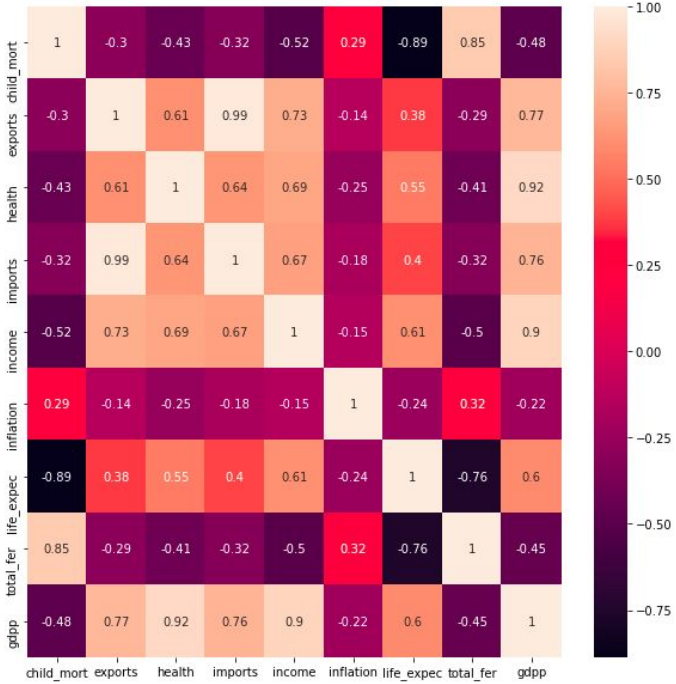

0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0 20000 40000 60000 80000 100000 120000
gdpp


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 20 40 60 80 100
inflation


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0 2 4 6 8
total_fer


0.05
0.04
0.03
0.02
0.01
0.00
20 30 40 50 60 70 80 90
life_expec

Ploting Heatmap for the numerical variables



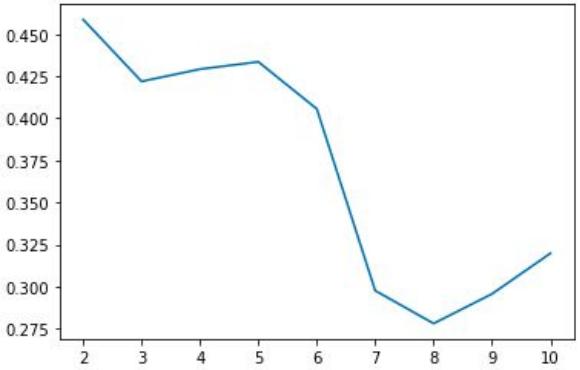
Inferences:

- We can see that health and gdp has highest correlation, followed by gdpp & exports.
- The least correlation is between life_expec & child_mort, followed by total_fer & life_expec

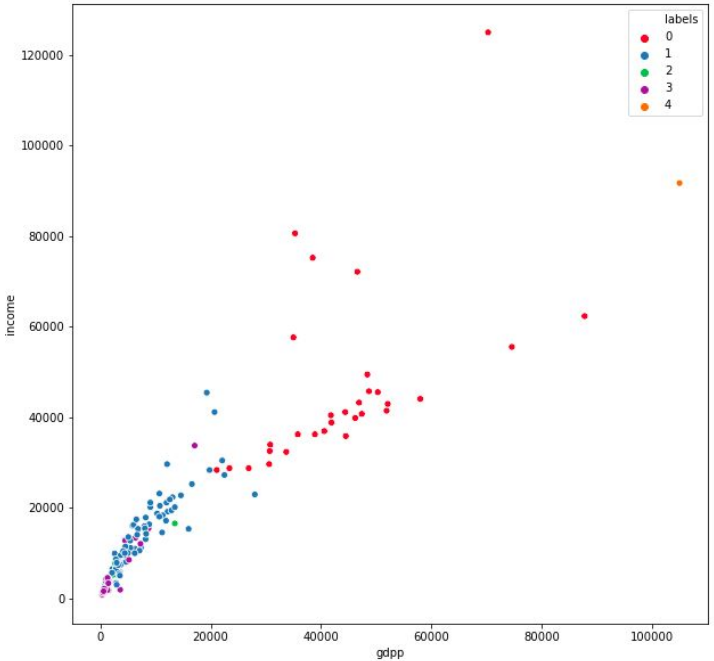
Silhoutte Score graph after scaling the data

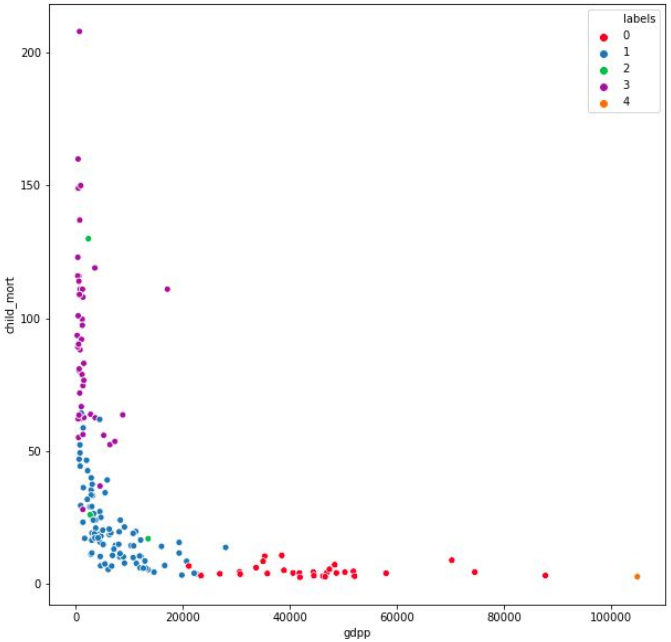
- From this we can conclude that highest value is for k=5
- Followed by k=4 and k=3

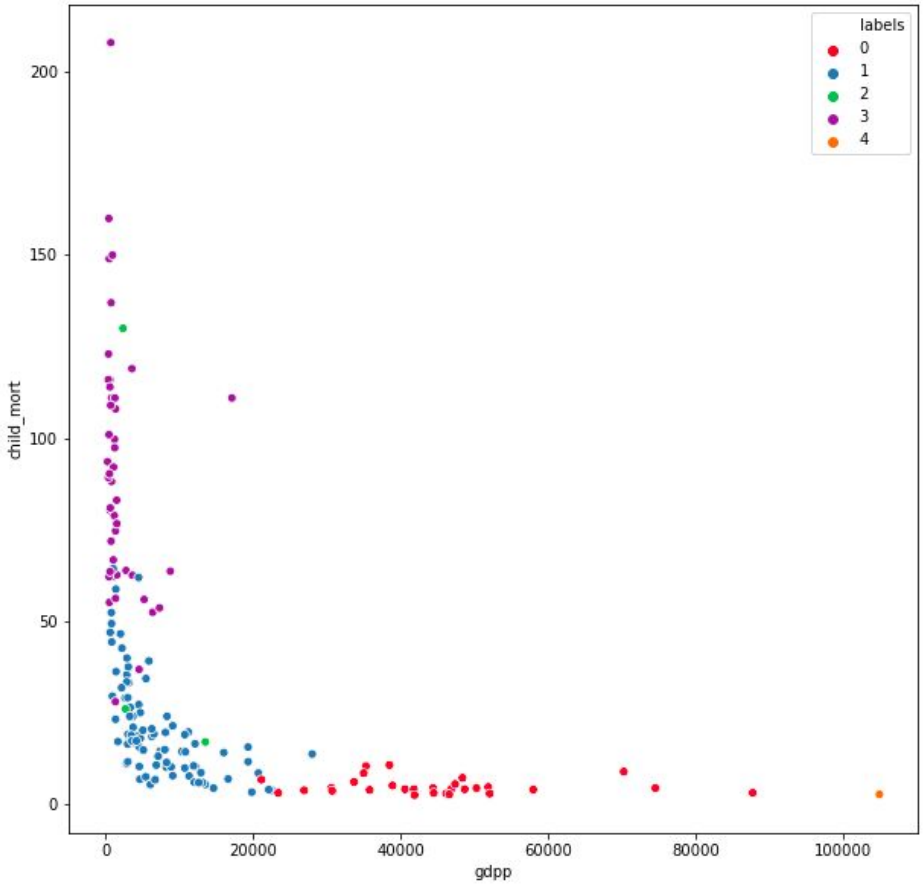
	0	1
0	2	0.458633
1	3	0.421862
2	4	0.429147
3	5	0.433475
4	6	0.405524
5	7	0.297558
6	8	0.278032
7	9	0.295637
8	10	0.319739



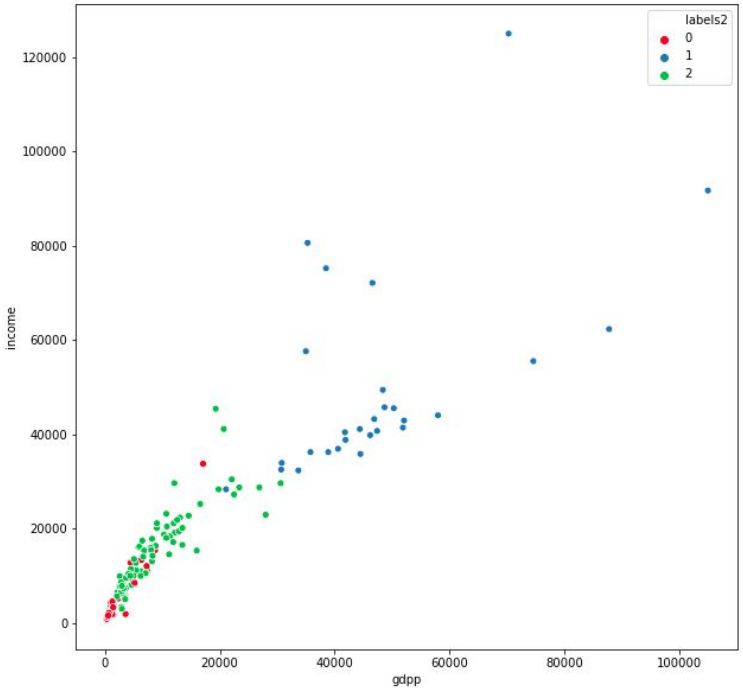
Perform analysis for k=5

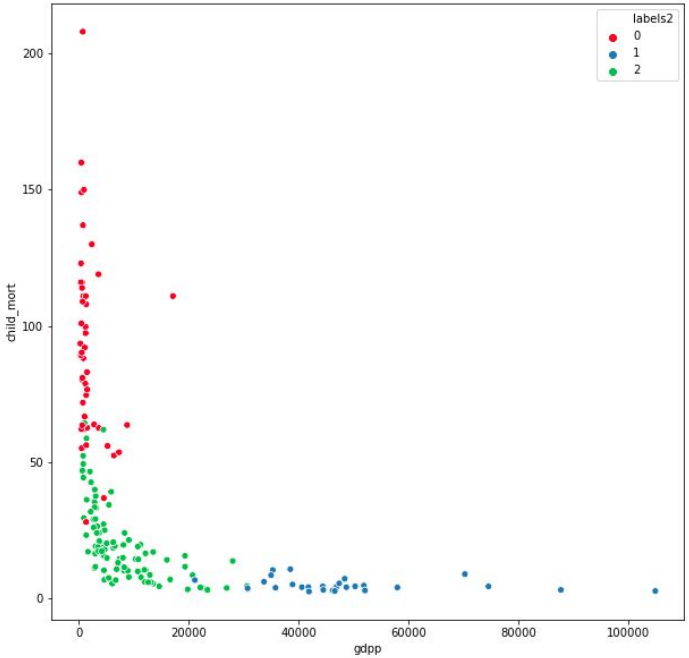


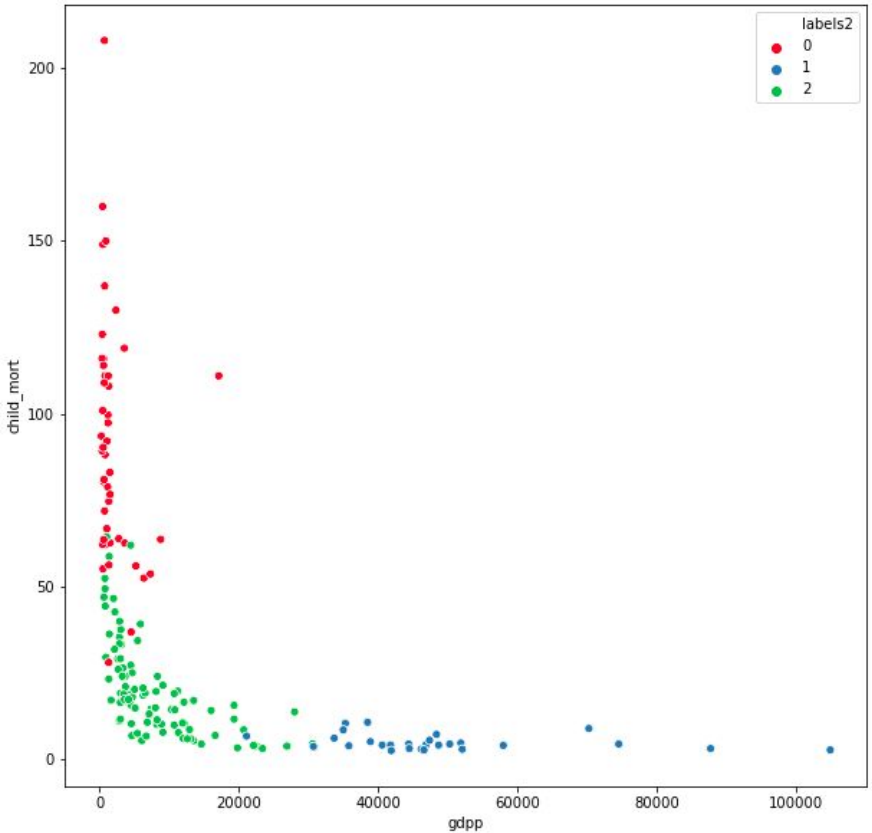



labels
0
1
2
3
4
child_mort
gdpp
0
50
100
150
200
0
20000
40000
60000
80000
100000

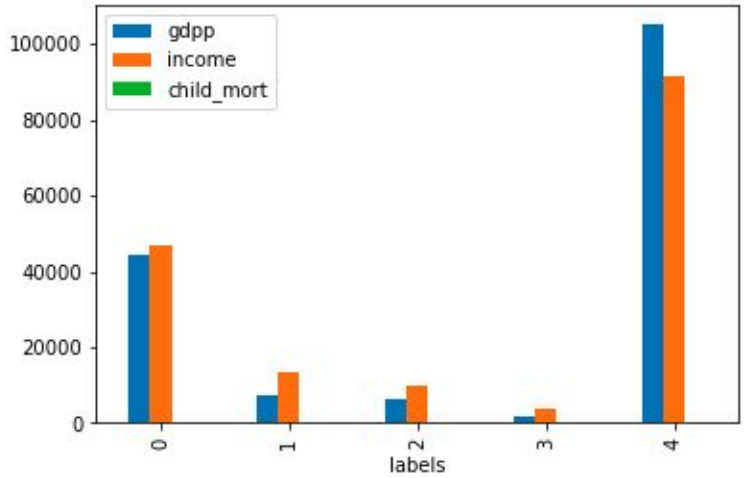
Analysis when k=3

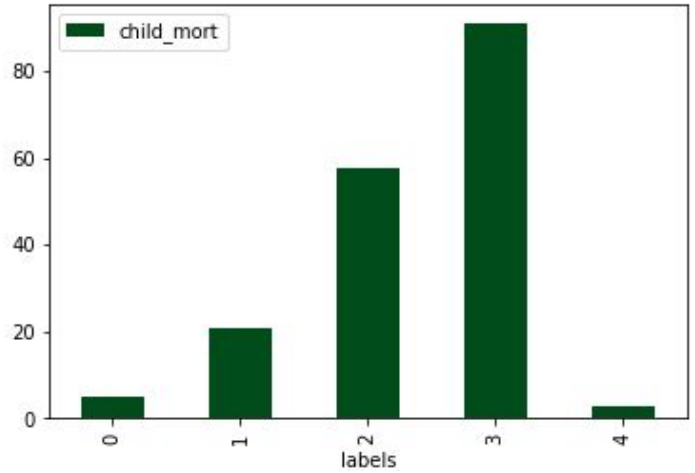




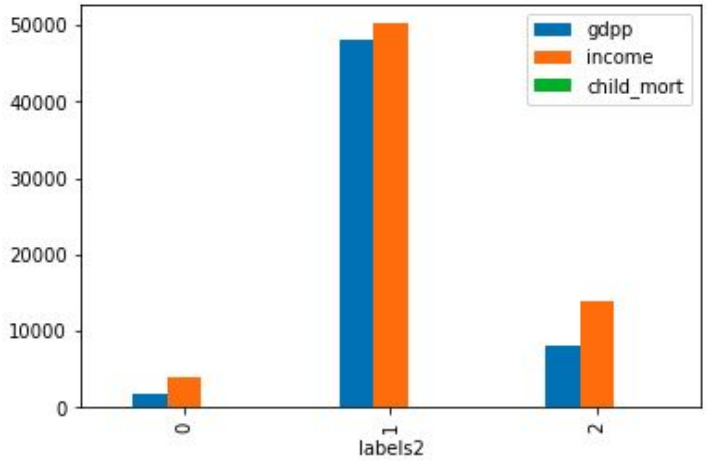

labels2
0
1
2
child_mort
gdpp
200
150
100
50
0
0
20000
40000
60000
80000
100000

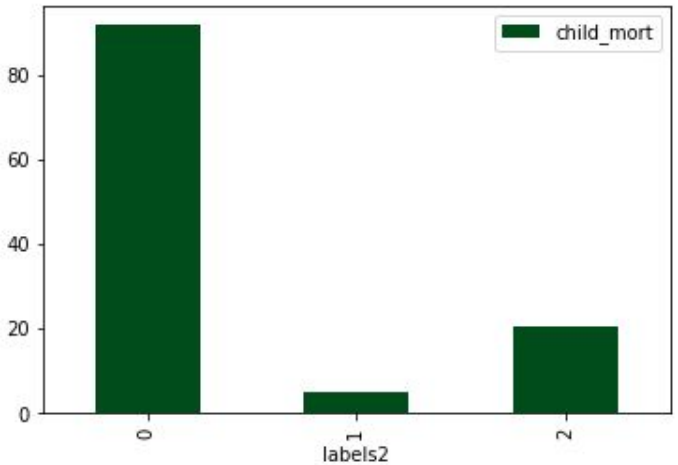
Perfoming profile clustering for k=5





Perfoming profile clustering for k=3

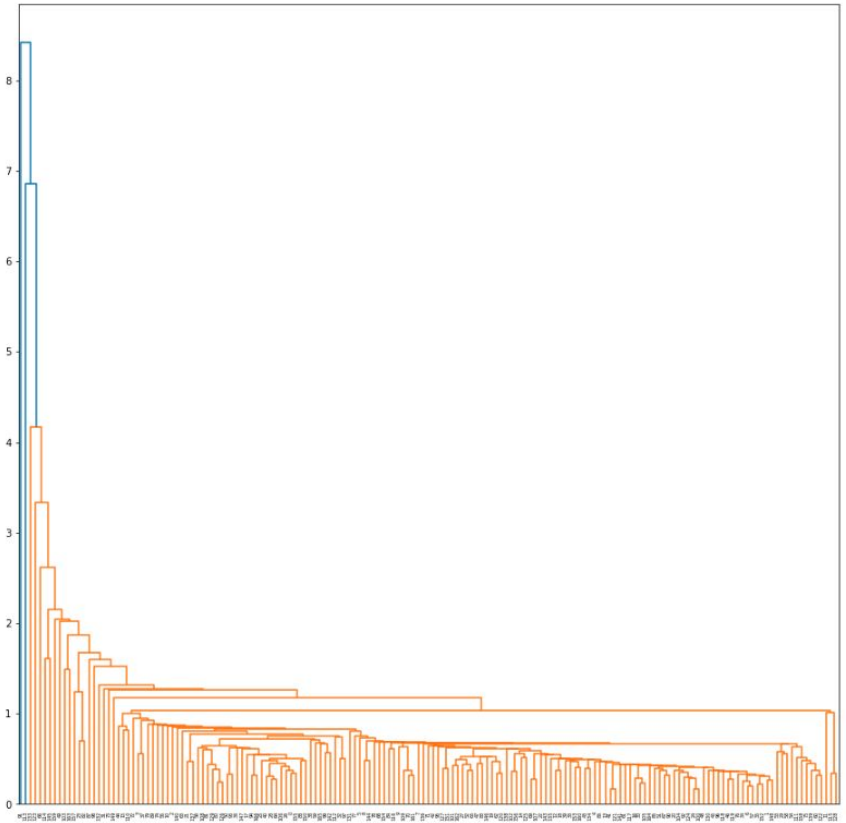




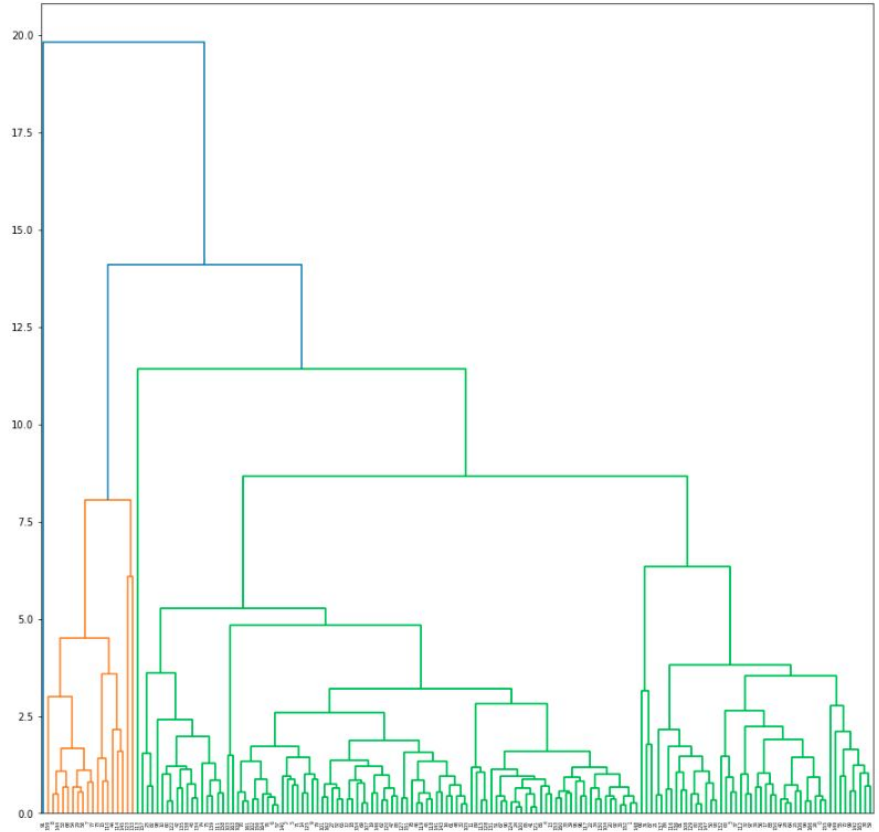
Inferences

- For the number of clusters=3, in cluster 0 we can see that child mortality is high, gdpp and income is very less
- So we can select the cluster 0 for our final reference

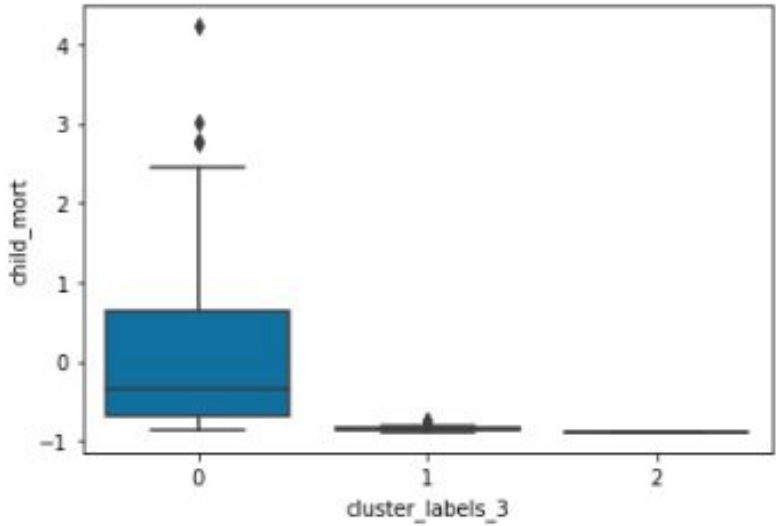
Single Linkage Hierarchical clustering



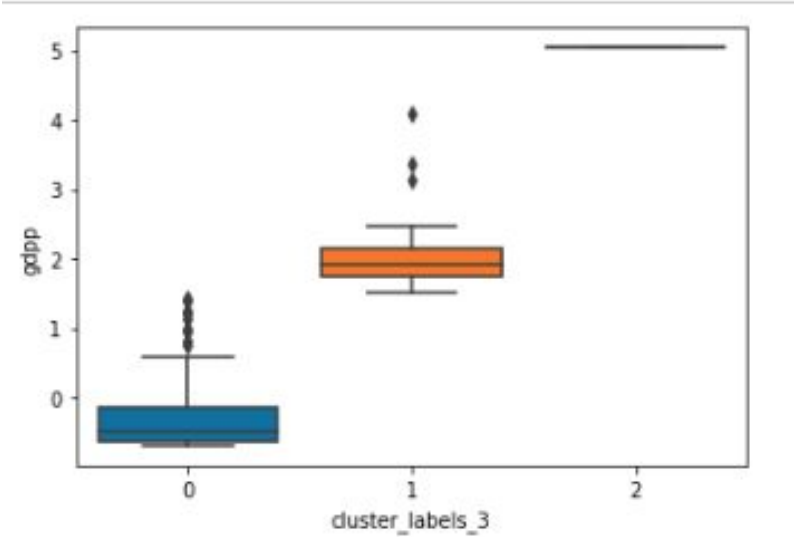
Complete Linkage Hierarchical clustering



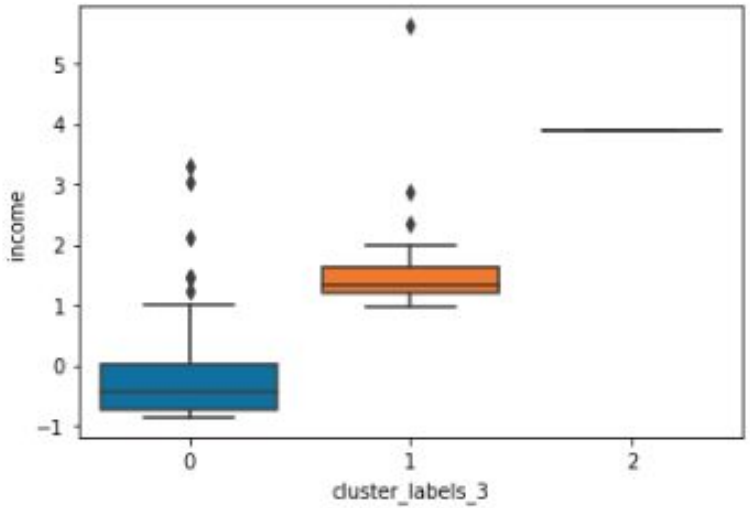
Box plot analysis for clusterlabel=3 vs child_mort



Box plot analysis for clusterlabel=3 vs gdpp



Box plot analysis for clusterlabel=3 vs income



Final Conclusion

The top 5 countries that are dire need of help based on hign child_mort, low income, low gdpp are:

1. Solomon Islands
2. Iraq
3. Botswana
4. South Africa
5. Eritrea